Estimation of Semiparametric Models in the Presence of Endogeneity and Sample Selection

Siddhartha Chib* Edward Greenberg Ivan Jeliazkov

September 10, 2008

Abstract
We analyze a semiparametric model for data that suffer from the problems of sample selection, where some of the data are observed for only part of the sample with a probability that depends on a selection equation, and of endogeneity, where a covariate is correlated with the disturbance term. The introduction of nonparametric functions in the model permits great flexibility in the way covariates affect response variables. We present an efficient Bayesian method for the analysis of such models that allows us to consider general systems of outcome variables and endogenous regressors that are continuous, binary, censored, or ordered. Estimation is by Markov chain Monte Carlo (MCMC) methods. The algorithm we propose does not require simulation of the outcomes that are missing due to the selection mechanism, which reduces the computational load and improves the mixing of the MCMC chain. The approach is applied to a model of women’s labor force participation and log-wage determination.

Keywords: Binary data; censored regression; data augmentation; incidental truncation; informative missingness; labor force participation; log-wage estimation; Markov chain Monte Carlo; model selection; sample selection; Tobit regression.

1 Introduction

In this article we extend the standard regression model by simultaneously allowing for sample selection, endogeneity, and nonparametric covariate effects. While each of these issues by itself can lead to complications in estimation, their joint presence brings out additional estimation challenges that require careful analysis.

The problem of sample selection arises when all variables are observed for a subset of the observational units—the selected sample—but only some of the variables are observed for the entire set—the potential sample. The factors that determine membership in the selected sample are often correlated with those that determine the outcome. This sample selection

---

* Siddhartha Chib is Harry C. Hartkopf Professor of Econometrics and Statistics, Olin Business School, Washington University in St. Louis; e-mail: chib@wustl.edu. Edward Greenberg is Professor Emeritus of Economics, Washington University in St. Louis; e-mail: edg@artsci.wustl.edu. Ivan Jeliazkov is Assistant Professor of Economics, University of California, Irvine; e-mail: ivan@uci.edu. We thank the Weidenbaum Center on the Economy, Government, and Public Policy for financial assistance.
problem is sometimes referred to as “non-ignorable truncation,” “incidental truncation,” or “informative missingness.”

We allow for endogenous covariates because such covariates are often the norm in observational studies. A well-known example of endogeneity occurs in the estimation of the effect of education on wages. In this case, it is possible that education may depend on such unobserved factors as ability and motivation that also have an effect on wages. Education is then an endogenous covariate and its effect on wages is confounded with the effect of ability and motivation on wages.

Finally, our model contains nonparametric covariate effects because the negative consequences of undetected nonlinearity on the estimation of covariate effects can be severe. If, for example, the effect of an endogenous covariate is mistakenly modeled linearly or as a low-order polynomial, the conditional distribution of the responses and the joint distribution of the errors may appear to be non-Gaussian even if the true data generating process is Gaussian. Misspecified nonlinearity can lead to misleading estimates for all covariate effects because misspecification in one equation can affect estimates of parameters, functions, and error distributions in other equations.

To illustrate a setting to which our model can be applied, consider data collected on a sample of married women who supply information about their hours of work, wage rates, on-the-job training, years of formal education, and such exogenous covariates as number and ages of children, age, and experience. This group is the potential sample. An important feature of these data is that information on wage rates and on on-the-job training is available only for women reporting positive hours of work—these observations constitute the selected sample. The primary response variable of interest is log-wage rates, but on-the-job training and formal education are endogenous because they may be correlated with such unobservable determinants of log-wages as ability or motivation. Finally, there could be reason to believe that such covariates as age and years of education affect wages nonlinearly.

In early work, Heckman (1976, 1979) devised a two-step estimation procedure for a prototypical sample selection model. Many variants of the basic model are summarized in Wooldridge (2002) together with a number of alternatives to the two-step procedure; a comparison of several of these alternatives is given in Puhani (2000). The standard approach for confronting endogeneity from both the frequentist and Bayesian viewpoints is through
the use of instrumental variables, which are variables that are uncorrelated with the error in the response variable and correlated with the endogenous covariate. For an extensive summary of these ideas, see Wooldridge (2002, chapters 5 and 17).

There is considerable Bayesian and frequentist work on relaxing the assumption that covariate effects are parametric. The underlying ideas can be traced to Whittaker (1923) and are summarized in Wahba (1978), Silverman (1985), Hastie and Tibshirani (1990), Denison, Holmes, Mallick, and Smith (2002), and Wasserman (2006). Nonparametric techniques have been applied to multi-equation systems with exogenous covariates such as the seemingly unrelated regression model in Smith and Kohn (2000), Holmes, Denison, and Mallick (2002), and Koop, Poirier, and Tobias (2005). These papers focus on estimation and model choice in a multi-equation setting, which may include models for interaction between covariates, and we extend those techniques to account for sample selection issues and endogeneity. Recent work that allows for nonparametric functions with endogenous variables, but without the complication of informative missingness, includes Chib and Greenberg (2007) and Hall and Horowitz (2005). Gallant and Nychka (1987) discuss semi-nonparametric estimators for the basic sample selection model without endogenous covariates. The semi-nonparametric estimators are series expansions, where the order of the expansion is allowed to increase as the sample size grows. For a given sample size, this procedure is equivalent to approximating the covariate function by a polynomial. Similar series expansion estimators are considered in Das, Newey, and Vella (2003) for a model with endogenous covariates and sample selection. They present two- and three-step estimators in the spirit of Heckman’s procedure to study the impact of education on wages, but the model utilizes a linear probability selection mechanism that does not constrain the probability of selection to be within $[0,1]$. In addition, in finite samples the series expansion only spans the class of low-order polynomials and the setup precludes the possibility of inference on how much smoothing is desirable because conventional optimal smoothing results do not apply. Moreover, they do not develop asymptotic results for the case in which some of the endogenous covariates are incidentally truncated. In contrast, we present a fully Bayesian, finite-sample inferential framework that accommodates endogeneity, sample selection, and flexible nonlinear effects within a formal probabilistic sample selection mechanism.

In Section 2, we present a hierarchical Bayesian model that accommodates the three
components of the regression model discussed above. In Section 3, we present easily implemented simulation methods to fit the model, and, in Section 4, we address the problem of model choice by discussing the computation of marginal likelihoods and Bayes factors to determine the posterior probabilities of competing models. We report on the performance of our techniques in a simulation study in Section 5 and analyze an application dealing with the labor supply of married women in Section 6. Section 7 offers concluding remarks.

2 The Model

Our model contains equations for a set of \( J = J_1 + J_2 \) variables, of which \( J_1 \) are observed only in the selected sample and \( J_2 \) other variables, including the selection variable, are always observed. Some of the variables may be endogenous covariates, which may be observed only in the selected sample or for all units in the sample. To reduce notational complexity we describe the model and estimation methodology in detail for \( J = 4 \) variables \( (y_1, \ldots, y_4) \), of which two \( (y_1 \text{ and } y_2) \) are observed only in the selected sample and two \( (y_3 \text{ and } y_4) \) are always observed. The response variable \( y_1 \) is the primary variable of interest. The variables \( y_2 \) and \( y_3 \) are endogenous regressors in the model for the primary response, and \( y_4 \) is a Tobit (or censored) selection variable that is either zero or positive (see Tobin, 1958). The case of a binary selection variable is discussed in Section 3.2. The exogenous covariates, denoted by \( w \) and \( x \), are assumed to be observed whenever the corresponding response variables they determine are observed. Special cases of the model that allow either or both of \( y_2 \) and \( y_3 \) to be absent do not require conceptual changes to the estimation algorithm, and generalizations to more variables are straightforward provided they do not lead to multicollinearity or the related problem of concurvity in nonparametric additive regression (Hastie and Tibshirani, 1990).

In detail, for subject \( i = 1, \ldots, n \), the model we analyze is

\[
\begin{align*}
  y_{1i} &= x_{i1}' \beta_1 + g_1(y_{i2}, y_{i3}, w_{i1}) + \varepsilon_{i1}, \quad (1) \\
  y_{2i} &= x_{i2}' \beta_2 + g_2(w_{i2}) + \varepsilon_{i2}, \quad (2) \\
  y_{3i} &= x_{i3}' \beta_3 + g_3(w_{i3}) + \varepsilon_{i3}, \quad (3) \\
  y_{4i}^* &= x_{i4}' \beta_4 + g_4(w_{i4}) + \varepsilon_{i4}, \quad (4)
\end{align*}
\]

where the first equation models the primary response of interest \( y_{1i} \), the second and third
equations model the endogenous regressors \( y_{i2} \) and \( y_{i3} \), and the fourth equation is the model for the latent censored selection variable \( y_{i4}^* \). The latent \( y_{i4}^* \) is related to the observed selection variable \( y_{i4} \) by \( y_{i4} = y_{i4}^* I(y_{i4}^* > 0) \), where \( I(\cdot) \) is the indicator function. The selection variable \( y_{i4} \) determines the set of variables that are observed for the \( i \)th unit in the sample: if \( y_{i4} > 0 \), the entire vector \( y_{i1:4} = (y_{i1}, y_{i2}, y_{i3}, y_{i4})' \) is observed, and when \( y_{i4} = 0 \), only \( y_{i3:4} = (y_{i3}, y_{i4})' \) are observed, and \( y_{i1:2} = (y_{i1}, y_{i2})' \) are missing. As an illustration, recall the married women’s wage data example mentioned in the introduction, where we let \( y_{i1} \) be the log-wage rate, \( y_{i2} \) be on-the-job training, \( y_{i3} \) be years of formal education, and \( y_{i4} \) be hours of work. The vectors \( x_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}) \) and \( w_i = (w_{i1}, w_{i2}, w_{i3}, w_{i4}) \) are exogenous covariates, where the effects of \( x_{ij} \) are linear and those of \( w_{ij} \) are nonparametric. An important feature of this model is that the effects of the endogenous variables \( y_{i2} \) and \( y_{i3} \) on the primary response \( y_{i1} \) may be nonparametric, although either or both can be entered linearly if desired. Correlation among \( (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3}) \) causes \( y_{i2} \) and \( y_{i3} \) to be correlated with the error term \( \varepsilon_{i1} \) in that equation (endogeneity), and correlation between \( \varepsilon_{i1} \) and \( \varepsilon_{i4} \) implies that \( y_{i4} \) is informative about \( \varepsilon_{i1} \) (sample selection). As a result, endogeneity and sample selection must be formally taken into account in the modeling and estimation.

To model the unknown functions in the \( j \)th equation, we assume the additive nonparametric structure discussed, for example, by Hastie and Tibshirani (1990),

\[
g_j(s_j) = \sum_{k=1}^{q_j} g_{jk}(s_{jk}),
\]

where \( s_{jk} \) is the \( k \)th covariate in \( s_j \) and \( q_j \) is the number of covariates in \( s_j \); the \( g_{jk}(\cdot) \) are nonparametric functions described below. The additive formulation is convenient because the “curse of dimensionality” renders nonparametric estimation of high-dimensional surfaces infeasible at present. Additive models are simple, easily interpretable, and sufficiently flexible for many practical applications. Additional flexibility can be attained by including covariate interactions in the set of regressors modeled additively, as long as doing so does not lead to concurvity.

For identification reasons we assume that the covariates in \( (x_2, w_2) \) and \( (x_3, w_3) \) contain at least one more variable than those included in \( (x_1, w_1) \). These variables can be regarded as instrumental variables. Although identification in models with incidental truncation does not require instruments, we assume in our applications that \( (x_4, w_4) \) contains covariates in
addition to those in \((x_1, w_1)\). Several variants of this model can be specified: the selection variable can be binary (e.g., labor market status) rather than censored (e.g., hours of work), and the remaining endogenous variables may be censored, ordered, or binary. We explain below how our methods can be applied to such cases. It is also straightforward to allow \(y_2\) and \(y_3\) to be vectors of endogenous variables.

The model is completed by assuming that the errors \(\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3}, \varepsilon_{i4})'\) have a multivariate normal distribution \(N_4(0, \Omega)\), where \(\Omega\) is an unrestricted symmetric positive definite matrix. It is possible to entertain other distributional forms for this joint distribution, but the normal assumption is important because it provides the underpinning for more flexible distributions, such as finite mixture distributions or continuous scale mixture distributions.

A final point is that a normality assumption in conjunction with non-parametric functions is much more flexible than it may seem. For example, in the case of a binary selection mechanism, \(P(y_{i4} = 1 | g_4) = F(g_4(w_{i4}))\), the marginal probit selection mechanism is fully flexible because \(g_4(\cdot)\) is unrestricted, even though the link function \(F(\cdot)\) is the c.d.f. of the Gaussian distribution. This way of modeling has advantages over modeling the distribution of the errors flexibly but considering only parametric effects in the mean of the selection equation; in the latter model, the effect of the \(x_4\) covariates is monotonic since \(F\) is monotonic and the mean is linear, but this is not necessarily the case when \(g_4(\cdot)\) is nonparametric.

### 2.1 The Likelihood Function

We begin the development of our algorithm with a discussion of the likelihood function of model (1)–(4). For the computations that follow, we define the vectors

\[
\mathbf{y}_{i3:4}^* = (y_{i3}, y_{i4}^*)', \quad \mathbf{y}_{i1:4}^* = (y_{i1}, y_{i2}, y_{i3}, y_{i4}^*),
\]

\[
\mathbf{g}_{i1:2} = (g_1(y_{i2}, y_{i3}, w_{i1}), g_2(w_{i2}))', \quad \mathbf{g}_{i3:4} = (g_3(w_{i3}), g_4(w_{i4}))', \quad \mathbf{g}_{i1:4} = (\mathbf{g}_{i1:2}', \mathbf{g}_{i3:4}')',
\]

\[
\mathbf{X}_{i3:4} = \begin{pmatrix} x_{i3}' & 0 & x_{i4}' \\ 0 & x_{i4}' \end{pmatrix}, \quad \mathbf{X}_{i1:4} = \begin{pmatrix} x_{i1}' & 0 & 0 & 0 \\ 0 & x_{i2}' & 0 & 0 \\ 0 & 0 & x_{i3}' & 0 \\ 0 & 0 & 0 & x_{i4}' \end{pmatrix}.
\]

We also set \(N_1 = \{i : y_{i4} > 0\}\) to be the \(n_1\) observations in the selected sample and \(N_2 = \{i : y_{i4} = 0\}\) to be the \(n_2\) observations in the potential sample that are not in the
selected sample. These definitions imply that $N_1$ is the set of indices for which we observe all four variables in $y_{i1:4}$, and $N_2$ is the set for which we observe only $y_{i3:4}$. Finally, let $\theta$ be the set of all model parameters and nonparametric functions.

The complete-data density function of the observations and latent data conditioned on $\theta$ is given by

$$f(y, y^*_4 | \theta) = \prod_{i \in N_1} f(y_{i1:4} | \theta) \prod_{i \in N_2} f(y^*_{i3:4} | \theta) I(y^*_4 < 0),$$

where $y$ contains $y_{i1:4}$ for $i \in N_1$ and $y_{i3:4}$ for $i \notin N_2$, while $y^*_4$ contains the observations on $y_{i4}^*$ for $i \notin N_2$. The second product on the right-hand side of equation (6) is derived from

$$f(y_{i3}, y_{i4}, y_{i4}^* | \theta) = f(y_{i3}, y_{i4}^* | \theta) \Pr(y_{i4} = 0 | y_{i3}, y_{i4}^*, \theta),$$

because $y_{i4} = 0$ for $i \notin N_2$, and $\Pr(y_{i4} = 0 | y_{i3}, y_{i4}^*, \theta) = I(y_{i4}^* < 0)$. Now partition $\Omega$ as

$$\Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix}^{(2 \times 2)},$$

upon defining $\beta = (\beta'_1, \beta'_2, \beta'_3, \beta'_4)'$ and

$$J = \begin{pmatrix} 0 \\ I \end{pmatrix},$$

so that $J'\beta = (\beta'_3, \beta'_4)'$, we have, for $i \in N_1$,

$$f(y_{i1:4} | \theta) \propto |\Omega|^{-1/2} \exp \left\{ -\frac{1}{2} (y_{i1:4} - g_{i1:4} - X_{i1:4}\beta)'\Omega^{-1}(y_{i1:4} - g_{i1:4} - X_{i1:4}\beta) \right\},$$

and for $i \notin N_2$,

$$f(y^*_{i3:4} | \theta) \propto |\Omega_{22}|^{-1/2} \exp \left\{ -\frac{1}{2} (y^*_{i3:4} - g_{i3:4} - X_{i3:4}J'\beta)'\Omega_{22}^{-1}(y^*_{i3:4} - g_{i3:4} - X_{i3:4}J'\beta) \right\},$$

which gives us the terms needed in (6). The above decomposition of $f(y, y^*_4 | \theta)$ is the basis for the straightforward and efficient sampler that we utilize in the remainder of the paper, but some computations require the likelihood function $f(y | \theta)$ marginally of the latent $y^*_4$. For those cases, it is convenient to write the complete-data likelihood function of the observations and latent data as

$$f(y, y^*_4 | \theta) = \prod_{i \in N_1} f(y_{i2:3} | \theta) f(y_{i1} | y_{i2:3}, \theta) f(y_{i4} | y_{i1:3}, \theta) \prod_{i \in N_2} f(y_{i3} | \theta) f(y_{i4}^* | y_{i3}, \theta) I(y_{i4} < 0)$$

(9).
where all densities are Gaussian, even though jointly they are not. As is standard in
censored-data models, the likelihood function \( f(y|\theta) \) is obtained by integrating \( f(y,y^*_i|\theta) \)
over the latent data \( y^*_i \), which is easily accomplished for the decomposition in (9) because
for each \( i \in N_2 \) such integration only involves computing the c.d.f. of a univariate Gaussian
distribution.

### 2.2 Prior distributions

We complete the model by specifying the prior distributions for the parameters and the
nonparametric functions. We assume that \( \beta \) has a joint normal distribution with mean \( \beta_0 \)
and variance \( B_0 \) and (independently) that the covariance matrix \( \Omega \) has an inverted Wishart
distribution with parameters \( \nu \) and \( Q \),

\[
\pi(\beta, \Omega) = \mathcal{N}(\beta|\beta_0, B_0) \mathcal{IW}(\Omega|\nu, Q),
\]

where \( \mathcal{N}(\beta|\beta_0, B_0) \) is the density of the multivariate normal distribution and \( \mathcal{IW}(\Omega|\nu, Q) \)
is the density of the inverse Wishart distribution.

We model each of the unknown functions through the class of Markov process smooth-
ness priors. This prior is easy to interpret, can approximate unknown functions arbitrarily
well with a penalty for “rough” functions, and has been widely used; see, for example, Shiller
(1973, 1984), Gersovitz and MacKinnon (1978), Besag, Green, Higdon, and Mengersen
(1995), Fahrmeir and Tutz (1997, Chapter 8), Müller, Rosner, Inoue, and Dewhirst (2001),

We note that other nonparametric modeling approaches could be applied in our setting
without altering our way of dealing with sample selection and endogeneity issues. These
include the integrated Wiener process priors proposed in Wahba (1978) and applied within
Bayesian inference by Wood and Kohn (1998) and Shively, Kohn, and Wood (1999). Other
approaches include regression splines (Smith and Kohn 1996), B-spline priors (Silverman
1985), and wavelets (Denison, Holmes, Mallick, and Smith 2002). Recent extensions of these
techniques to Bayesian free-knot spline modeling are given in Denison, Mallick, and Smith
(1998) and DiMatteo, Genovese, and Kass (2001). Extensions to multivariate functions are
See Wasserman (2006) for a discussion of different nonparametric modeling approaches from
the frequentist viewpoint or Denison et al. (2002) for a Bayesian perspective.
Although alternative nonparametric modeling approaches can be pursued for the unknown functions, two issues that can arise when alternatives are considered should be kept in mind. First, computations involving some non-parametric approaches, e.g., the integrated Wiener process prior approach, may involve $O(n^3)$ operations, an important consideration for large samples. While adaptive approaches to knot selection (see Denison, et al. 1998 and DiMatteo et al. 2001) can reduce the computational burden, these approaches are also costly to implement because selecting the number and location of knots is non-trivial. Second, many classes of priors used in non-parametric functional modeling lead to partially improper priors on the unknown functions that are problematic for model choice. The approach we follow leads to an efficient estimation algorithm that requires $O(n)$ rather than $O(n^3)$ computations and produces proper priors for the unknown functions and parameters.

Because the unknown functions are treated similarly and assumed to be a priori independent, it is sufficient to give the details for any one of the univariate nonparametric functions in (5). To understand the main modeling issues in our setting, consider the $j$th equation, where interest focuses on the $k$th function $g_{jk}(\cdot)$ in that equation ($j = 1, \ldots, J$, $k = 1, \ldots, q_j$). If $j \leq J_1$ (i.e., equation $j$ is observed only in the selected sample), the vector of covariates $w_{jk}$ that enters $g_{jk}(\cdot)$ contains $n_1$ observations, i.e. $w_{jk} = (w_{1jk}, \ldots, w_{n_1jk})'$ consisting of $\{w_{ijk}\}$ for $i \in \mathbb{N}_1$; otherwise, when $J_1 < j \leq J$ (i.e., equation $j$ is always observed), we have $w_{jk} = (w_{1jk}, \ldots, w_{njk})'$.

Since there may be repeated values in $w_{jk}$, we define the $p_{jk} \times 1$ design point vector $v_{jk} = (v_{jk,1}, \ldots, v_{jk,p_{jk}})'$ of unique ordered values of $w_{jk}$ with $v_{jk,1} < \ldots < v_{jk,p_{jk}}$, where $p_{jk} \leq n_1$ if $j \leq J_1$ or $p_{jk} \leq n$ if $J_1 < j \leq J$. The notation used for the elements in $v_{jk}$ is intended to emphasize that enumeration in that vector does not correspond to the enumeration in $w_{jk}$, whose entries need neither to be ordered nor unique. Given the unique and ordered values $v_{jk}$ and defining $g_{jk,t} = g_{jk}(v_{jk,t})$, the basic idea is to model the vector of functional evaluations $\left(g_{jk,1}, \ldots, g_{jk,p_{jk}}\right) \equiv \left(g_{jk}(v_{jk,1}), \ldots, g_{jk}(v_{jk,p_{jk}})\right)'$ by viewing them as the realization of a stochastic process that both allow flexibility and also penalize sharp differences between successive functional evaluations.

Before continuing with details on the stochastic model for $\{g_{jk,t}\}$, we note that unrestricted additive models are identified only up to a constant because the likelihood remains unchanged if $g_{jk}(\cdot)$ and $g_{jh}(\cdot)$, $k \neq h$, in (5) are simultaneously redefined as $g_{jk}^*(\cdot) = g_{jk}(\cdot) + a$.
and \( g_{jh}^*(\cdot) = g_{jh}(\cdot) - a \) for some constant \( a \), so that \( g_{jk}(\cdot) + g_{jh}(\cdot) = g_{jk}^*(\cdot) + g_{jh}^*(\cdot) \). To achieve identification, the nonparametric functions must be restricted to remove any free constants. We follow the approach of Shively et al. (1999) by restricting the functions to equal zero at the first ordered observation (i.e., \( g_{jk,1} = 0 \)), which allows the parametric part of the model to absorb the overall intercept. For the prior distribution of the second state \( g_{jk,2} \) of the process, we assume

\[
g_{jk,2} | \tau_{jk}^2 \sim \mathcal{N} \left( g_{jk,0,2}, \tau_{jk}^2 \tau_{jk}^* G_{jk,0,2} \right),
\]

where \( \tau_{jk}^2 \) is a smoothness parameter discussed below.

We model the remaining function evaluations as resulting from the realization of a second-order Markov process. With \( h_{jk,t} = v_{jk,t} - v_{jk,t-1} \), the second-order Markov process prior for \( g_{jk,t} \), \( t = 3, \ldots, p_{jk} \), is

\[
g_{jk,t} = \left( 1 + \frac{h_{jk,t}}{h_{jk,t-1}} \right) g_{jk,t-1} - \frac{h_{jk,t}}{h_{jk,t-1}} g_{jk,t-2} + u_{jk,t}, \quad u_{jk,t} \sim \mathcal{N} \left( 0, \tau_{jk}^2 h_{jk,t} \right),
\]

where \( \tau_{jk}^2 \) acts as a smoothness parameter in the sense that small values produce smooth functions and large values allow the function to interpolate the data more closely. This prior assumes that the variance grows linearly with the distance \( h_{jk,t} \), a property satisfied by random walks, but other choices are possible (see e.g., Shiller 1984, Besag et al. 1995, and Fahrmeir and Lang 2001).

To see more clearly how this prior specification introduces smoothness, we note several properties of (11). First, the expected value of \( g_{jk,t} \) given \( g_{jk,t-1} \) and \( g_{jk,t-2} \) lies on a straight line that passes through \( g_{jk,t-1} \) and \( g_{jk,t-2} \). Hence, the notion of smoothness that the prior in (11) emphasizes is that of local linearity. The possibility of departure from local linearity arises from \( u_{jk,t} \), whose variance controls the degree to which deviations are acceptable. The modeling is desirable because, unlike other modeling approaches for the unknown functions, it assumes neither continuous functions nor derivatives.

We include the \( \left\{ \tau_{jk}^2 \right\} \) \( (j = 1, \ldots, J, \) and \( k = 1, \ldots, q_j) \) in the sampler with prior distributions of the inverse gamma form

\[
\tau_{jk}^2 \sim \mathcal{IG}(\nu_{jk0}/2, \delta_{jk0}/2).
\]

The prior specified by equations (10), (11), and (12) yields a proper and computationally convenient joint prior distribution for the \( (p_{jk} - 1) \)-vector of unrestricted function
evaluations \( g_{jk} = (g_{jk,2}, \ldots, g_{jk,p})' \). To see the last point, note that after defining

\[
H_{jk} = \begin{pmatrix}
-\frac{1}{h_{jk,2}} & \frac{h_{jk,2}}{\tau_{jk}^2} & \cdots & \frac{h_{jk,p}}{\tau_{jk}^2} \\
\frac{h_{jk,2}}{\tau_{jk}^2} & -\frac{1}{h_{jk,3}} & \cdots & \frac{h_{jk,p}}{\tau_{jk}^2} \\
\vdots & \ddots & \ddots & \vdots \\
\frac{h_{jk,2}}{\tau_{jk}^2} & \cdots & \frac{h_{jk,p}}{\tau_{jk}^2} & -\frac{1}{h_{jk,p}}
\end{pmatrix}
\]

and \( \Sigma_{jk} = \text{diag}(G_{jk0}, h_{jk,3}, \ldots, h_{jk,p}) \), the joint distribution of \( g_{jk | \tau_{jk}^2} \) is Gaussian,

\[
g_{jk | \tau_{jk}^2} \sim N \left( g_{jk0}, \tau_{jk}^2 K_{jk}^{-1} \right),
\]

where \( g_{jk0} = H_{jk}^{-1} \tilde{g}_{jk0}, \tilde{g}_{jk0} = (g_{jk0,2}, 0, \ldots, 0)' \) and the penalty matrix \( K_{jk} = H_{jk}' \Sigma_{jk}^{-1} H_{jk} \). It is straightforward to verify that since \( H_{jk} \) produces second-order differences when post-multiplied by a vector, its inverse constructs second-order sums. As a result, the vector \( H_{jk}^{-1} \tilde{g}_{jk0} \) can be constructed by iterating (11) in expectation, starting with \( g_{jk0,1} = 0 \) and \( g_{jk0,2} \) from (10). It is important to note that \( K_{jk} \) is banded and that manipulations of banded matrices require \( O(n) \) operations, rather than the usual \( O(n^3) \) for inversions or \( O(n^2) \) for vector multiplication. Bandedness yields important computational savings in implementing the estimation algorithm, which is an important advantage over modeling approaches for the unknown function that do not result in banded precision or variance matrices. Although bandedness of the precision matrix is a feature of such models as the B-spline approach of Silverman (1985) and the wavelet modeling of Denison et al. (2002), it is not a feature of other modeling approaches, e.g., the integrated Wiener process prior of Wahba (1978).

3 Estimation

In this section we first specify an algorithm for the censored selection variable and then turn to the modifications necessary for other data types.

3.1 Algorithm for censored selection variable

For notational convenience, we define \( \theta \equiv (\beta, \Omega, \{g_{jk}\}, \{\tau_{jk}^2\}) \), but otherwise continue to use the notation developed in the preceeding sections. The posterior distribution for \( \theta \) and
\({\{y_{i4}^*}\}_{i\in N_2}\) for the semiparametric model of (1)–(4) is given by

\[
\pi(\theta, \{y_{i4}^*\}_{i\in N_2} | y) \propto \left[ \prod_{i\in N_1} f(y_{i1:4} | \theta) \right] \left[ \prod_{i\in N_2} f(y_{i4:3}^* | \theta) I(y_{i4}^* < 0) \right] \mathcal{N}(\beta | \beta_0, B_0) \\
\times I(W(\Omega | \nu, Q) \prod_{j,k} \mathcal{N}(g_{jk} | g_{jk0}, \tau_{jk}^{-2} K_{jk}^{-1}) \mathcal{IG}(\tau_{jk}^2 | \nu_{jk0}/2, \delta_{jk0}/2)).
\tag{14}
\]

This posterior distribution can be sampled by Markov chain Monte Carlo (MCMC) methods; see Chib (2001) for a detailed survey. Letting \(\theta \backslash \theta_k\) represent the elements of \(\theta\) other than \(\theta_k\), we summarize the algorithm as follows:

**Algorithm 1: MCMC estimation of nonparametric incidental truncation model**

1. Sample \(\beta\) from the distribution \(\beta | y, y_4^*, \theta \backslash \beta\).
2. Sample \(\Omega\) from the distribution \(\Omega | y, y_4^*, \theta \backslash \Omega\) in a one-block, three-step procedure.
3. For \(j = 1, \ldots, J, k = 1, \ldots, q_j\), sample \(g_{jk}\) from the distribution \(g_{jk} | y, y_4^*, \theta \backslash g_{jk}\).
4. For \(j = 1, \ldots, J, k = 1, \ldots, q_j\), sample \(\tau_{jk}^2\) from the distribution \(\tau_{jk}^2 | y, g_{jk}\).
5. For \(i \in N_2\), sample \(y_{i4}^*\) from the distribution \(y_{i4}^* | y, \theta\).

It is important to note that we do not involve the missing \(\{y_{i1:2}\}\) for \(i \in N_2\) in this algorithm; that is, we do not augment the selected sample with the data that are not part of the selected sample. This may seem surprising because having the augmented “full” potential sample would reduce the model to a nonparametric seemingly unrelated regression model that could be processed along the lines of Chib and Greenberg (1995), Smith and Kohn (2000), or Holmes, Denison, and Mallick (2002). Specifically, if the missing outcomes are denoted by \(\{y_{i1:2}^\dagger\}\), sampling could proceed recursively by drawing from \(\theta | y, y_4^*, \{y_{i1:2}^\dagger\}\), \(y_{i4}^* | y, \theta, \{y_{i1:2}^\dagger\}\), and \(\{y_{i1:2}^\dagger\} | y, y_4^*, \theta\), which would be updated by a series of full-conditional draws. In contrast, our proposed approach is computationally easier and has three major advantages over the approach in which the missing data are part of the sampling. First, it reduces computational and storage demands because simulation of the missing \(\{y_{i1:2}^\dagger\}\) is not needed. Second, it significantly improves the mixing of the Markov chain as sampling...
is not conditional on \( \{y_{1:2}^{\dagger}\} \). In Section 5 we present evidence of the improved simulation performance of our method relative to output from a chain that includes \( \{y_{1:2}^{\dagger}\} \) and show that the inefficiency factors are much larger when the outcomes that are missing due to the selection mechanism are included. Our results are consistent with the results of Liu (1994) and Liu, Wong, and Kong (1994), who show that collapsed Gibbs samplers can provide improved simulation performance. These two advantages of the sampler can become important when the proportion of missing outcomes is high or when the number of parameters is large.

Third, augmenting the sampler with the missing \( \{y_{1:2}^{\dagger}\} \) is not straightforward if some of the covariates are missing when the corresponding responses are missing, which would require models for the missing covariates that are not necessary in our approach. It should be noted that, if needed, inference about the missing covariates can be conducted ex-post from the probability density function of the missing data conditioned on parameters. We now turn to the details of the sampler.

**Sampling \( \beta \)** The posterior distribution of equation (14) implies \( \beta | y^*, \theta \beta \sim N(b, B) \), where

\[
  b = B \left( B_0^{-1} b_0 + \sum_{i \in N_1} X_i' \Omega^{-1} (y_{i1:4}^* - g_{i1:4}) + \sum_{i \in N_2} JX_i' \Omega_{22}^{-1} (y_{i3:4}^* - g_{i3:4}) \right)
\]

\[
  B = \left( B_0^{-1} + \sum_{i \in N_1} X_i' \Omega^{-1} X_i + \sum_{i \in N_2} JX_i' \Omega_{22}^{-1} X_i J' \right)^{-1}
\]

where \( J \) was defined in (8). This step proceeds without the unobserved \( y_{1:2} \) by computing the conditional mean and covariance of \( (\beta_1', \beta_2')' \) from the observations in \( N_1 \) and the conditional mean and covariance of \( (\beta_3', \beta_4')' \) from the observations in both \( N_1 \) and \( N_2 \). The matrix \( J \) selects \( \beta_3 \) and \( \beta_4 \) from \( \beta \) to include in the computations the observations in \( N_2 \). The partitioning of \( \Omega \) is necessary because the observations in \( N_2 \) are modeled by the third and fourth equations only.

**Sampling \( \Omega \)** Our decision not to sample the missing \( y_{1:2} \) also has implications for the way in which \( \Omega \) is sampled. In particular, the conditional distribution \( \Omega | y, y^*_4, \theta \Omega \) is not inverse Wishart, because of the different forms of the complete data likelihood in \( N_1 \) and \( N_2 \). It is nevertheless possible to derive and sample the distributions \( \Omega_{22} | y, y^*_4, \theta \Omega_{22} \),
\( \Omega_{11.2} | y, y^*_4, \theta \backslash \Omega_{11.2} \), and \( B_{21} | y, y^*_4, \Omega_{11.2} \), where

\[
\begin{align*}
\Omega_{11.2} &= \Omega_{11} - \Omega_{12} \Omega_{22}^{-1} \Omega_{21}, \\
B_{21} &= \Omega_{22}^{-1} \Omega_{21},
\end{align*}
\]

from which \( \Omega \) can be recovered. To see the form of these three conditional distributions, let

\[
\begin{align*}
\eta_{1:4} &= y_{1:4} - g_{1:4} - X_{1:4} \beta, \quad i \in N_1 \\
\eta^*_{3:4} &= y^*_{3:4} - g_{3:4} - X_{3:4} \beta', \quad i \in N_2.
\end{align*}
\]

The complete data likelihood is then

\[
\prod_{i \in N_1} f(\eta_{1:4} | \theta) \prod_{i \in N_2} f(\eta^*_{3:4} | \theta) \propto |\Omega|^{-n_{1:4}/2} \exp \left[ -\frac{1}{2} \sum_{i \in N_1} \eta_{1:4}' \Omega^{-1} \eta_{1:4} \right] \\
\times |\Omega_{22}|^{-n_{2:4}/2} \exp \left[ -\frac{1}{2} \sum_{i \in N_2} \eta^*_{3:4}' \Omega^{-1} \eta^*_{3:4} \right].
\]

Partitioning the hyperparameter matrix \( Q \) from the inverse Wishart prior conformably with \( \Omega \) as

\[
Q = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix},
\]

we find the posterior full-conditional distribution of \( \Omega \):

\[
\pi(\Omega | y, y^*_4, \theta \backslash \Omega) \propto |\Omega|^{-(\nu+n_1+J_1+J_2+1)/2} \exp \left[ -\frac{1}{2} \text{tr} (\Omega^{-1} R) \right] \\
\times |\Omega_{22}|^{-n_{2:4}/2} \exp \left[ -\frac{1}{2} \text{tr} \left( \Omega_{22}^{-1} \sum_{i \in N_2} \eta^*_{i:4} \eta^*_{i:4}' \right) \right],
\]

where \( R = Q + \sum_{i \in N_1} \eta_{i:4}' \eta_{i:4} \).

Making the change of variables from \( \Omega \) to \( (\Omega_{22}, \Omega_{11.2}, B_{21}) \), with Jacobian \( |\Omega_{22}|^{J_2} \), we obtain the posterior distribution

\[
\pi(\Omega_{22}, \Omega_{11.2}, B_{21} | y, y^*_4, \theta \backslash \Omega) \propto |\Omega_{11.2}|^{-(\nu+n_1+J_1+J_2+1)/2} \exp \left[ -\frac{1}{2} \text{tr} (\Omega^{-1} R) \right] \\
\times |\Omega_{22}|^{-(\nu-n_1-J_1-J_2+1)/2} \exp \left[ -\frac{1}{2} \text{tr} \left( \Omega_{22}^{-1} \sum_{i \in N_2} \eta^*_{i:4} \eta^*_{i:4}' \right) \right],
\]

where we use \( |\Omega| = |\Omega_{11.2}| |\Omega_{22}|. \) By the partitioned inverse theorem

\[
\Omega^{-1} = \begin{pmatrix} \Omega_{11.2}^{-1} & -\Omega_{11.2}^{-1} B_{21}^* \\ -B_{21} \Omega_{11.2}^{-1} & \Omega_{22}^{-1} + B_{21} \Omega_{11.2}^{-1} B_{21}^* \end{pmatrix},
\]

14
we are able to simplify the trace as

\[
\text{tr}(\Omega^{-1}R) = \text{tr}
\left[
\begin{align*}
\Omega_{11:2}^{-1} & R_{11} - \Omega_{11:2}^{-1} B_{21}' R_{21} - B_{21} \Omega_{11:2}^{-1} R_{12} \\
+ (\Omega_{22}^{-1} + B_{21} \Omega_{11:2}^{-1} B_{21}') R_{22}
\end{align*}
\right]
\]

\[
= \text{tr}
\left[
\begin{align*}
\Omega_{11:2}^{-1} & R_{11} + B_{21}' R_{22} B_{21} - B_{21} R_{21} - R_{12} B_{21}
\end{align*}
\right] + \text{tr}(\Omega_{22}^{-1} R_{22})
\]

where \( R \) has been partitioned to conform to \( Q \). It now follows that

\[
\pi(\Omega_{22}, \Omega_{11:2}, B_{21} | y^*, \theta \setminus \Omega) \propto |\Omega_{11:2}|^{-(\nu+n_1+J_1+1)/2} \exp \left[ -\frac{1}{2} \text{tr}(\Omega_{11:2}^{-1} R_{11:2}) \right]
\]

\[
\times |\Omega_{22}|^{-(\nu+n-J_2+1)/2} \exp \left[ -\frac{1}{2} \text{tr}(\Omega_{22}^{-1} (B_{21} - R_{22}^{-1} R_{22})' R_{22} (B_{21} - R_{22}^{-1} R_{22})) \right]
\]

\[
\times |\Omega_{11:2}|^{-(\nu+n-J_2+1)/2} \exp \left[ -\frac{1}{2} \text{tr}(\Omega_{22}^{-1} [R_{22} + \sum_{N_2} \eta_{3:4} \eta_{3:4}^*]) \right],
\]

where \( R_{11:2} = R_{11} - R_{12} R_{22}^{-1} R_{21} \). From here we conclude that

\[
\Omega_{22} | y, y^*, \theta \setminus \Omega \sim \mathcal{IW}(\nu - J_1 + n_1 + n_2, Q_{22} + \sum_{N_1, N_2} \eta_{3:4}^* \eta_{3:4}^*)
\]

\[
\Omega_{11:2} | y, y^*, \theta \setminus \Omega \sim \mathcal{IW}(\nu + n_1, R_{11:2})
\]

\[
B_{21} | y, y^*, \Omega_{11:2} \sim \mathcal{MN}_{J_2 \times J_1} (R_{22}^{-1} R_{21}, \Omega_{11:2} \otimes R_{22}^{-1}).
\]

The sampling can thus proceed from the full conditional densities of \( \Omega_{22}, \Omega_{11:2}, \) and \( B_{21}, \) from which \( \Omega \) can be recovered.

**Sampling the nonparametric functions** We sample the nonparametric functions one at a time, conditional on all remaining functions, parameters, and the latent data, by exploiting efficient \( O(n) \) sampling algorithms that utilize banded matrix operations. Because the sampling of each function is conditional on all parameters and other functions, in sampling the \( k \)th function we can focus only on the equation containing that function. Letting

\[
y_{ij}^* = \begin{cases} y_{ij}, & \text{if } j = 1, 2, 3, \\ y_{41}^*, & \text{if } j = 4, \end{cases}
\]

in order to isolate the \( k \)th function in equation \( j \), we define \( \xi_{ijk} \equiv y_{ij}^* - x_{ij}^* \beta_j - \sum_{h \neq k} g_{jh}(w_{ikh}) - E(\epsilon_{ij} | \xi_{i \setminus j}) \), so that

\[
\xi_{ijk} = g_{jk}(w_{ijk}) + \hat{\epsilon}_{ijk},
\]

\( 15 \)
where \( \hat{\varepsilon}_{ijk} \overset{\text{ind}}{\sim} N(0, \text{Var}(\varepsilon|\varepsilon_{i\setminus j})) \). Note that we can condition on \( \varepsilon_{i\setminus j} \) because we are conditioning upon \( \theta_{\setminus g} \) in this computation. Stacking over the \( n_1 \) observations if equation \( j \) is part of the selected sample (\( j \leq J_1 \)) or over the \( n \) observations if equation \( j \) is always observed (\( J_1 < j \leq J \)), we can write

\[
\xi_{jk} = P_{jk} g_{jk} + \hat{\varepsilon}_{jk},
\]

where \( g_{jk} \) is the \((p_{jk} - 1)\)-vector of unrestricted function evaluations defined on \( \tilde{v}_{jk} = (v_{jk,2}, \ldots, v_{jk,p_{jk}}) \) as discussed in Section 2.2. The matrix \( P_{jk} \) is an \( n \times (p_{jk} - 1) \) or \( n_1 \times (p_{jk} - 1) \) incidence matrix with entries \( P_{jk}(h, l) = 1 \) if \( w_{hjk} = v_{jk,l+1} \) and 0 otherwise, which establishes the correspondence between \( w_{jk} \) and \( \tilde{v}_{jk} \) over which the unrestricted function evaluations are defined. Since the rows of \( P_{jk} \) for which \( w_{hjk} = v_{jk,1} \) contain only zeros, whereas all other rows contain a single 1, row \( i \) of the product \( P_{jk} g_{jk} \) is \( g_{jk}(w_{ijk}) \).

A closer look at equation (15) can help demystify the nature of nonparametric modeling. It is instructive to note that, when written in this way, the model specifies a dummy variable at each unique covariate observation, with \( g_{jk}(v_{jk,1}) \) serving as the omitted category. In this model, the \( g_{jk} \) can be interpreted as the parameters on which the prior in Section 2.2 imposes smoothness without ruling out any values they may take, which explains why we consider \( g_{jk}(\cdot) \) to be a nonparametric function.

It now follows from standard calculations that

\[
g_{jk}|y^*, \theta_{\setminus g_{jk}} \sim N\left( \hat{g}_{jk}, \hat{G}_{jk} \right),
\]

where

\[
\hat{G}_{jk} = \left( \tau_{jk}^{-2} K_{jk} + P'_{jk} V^{-1}_{jk} P_{jk} \right)^{-1},
\]

\[
\hat{g}_{jk} = \hat{G}_{jk} \left( \tau_{jk}^{-2} K_{jk} g_{jk0} + P'_{jk} V^{-1}_{jk} \xi_{jk} \right),
\]

and \( V_{jk} \) is a diagonal matrix with entries equal to \( \text{Var}(\varepsilon_{ij}|\varepsilon_{i\setminus j}) \), which introduces heteroskedasticity into the sampling of the unknown functions.

In sampling \( g_{jk} \), one should note that \( P'_{jk} V^{-1}_{jk} P_{jk} \) is a diagonal matrix with the \( t \)-th diagonal entry equal to the number of values in \( w_{jk} \) corresponding to the \( t \)-th entry in \( \tilde{v}_{jk} \) divided by \( \text{Var}(\varepsilon_{ij}|\varepsilon_{i\setminus j}) \). Because \( K_{jk} \) and \( P'_{jk} V^{-1}_{jk} P_{jk} \) are banded, \( \hat{G}_{jk}^{-1} \) is banded as well, and sampling of \( g_{jk} \) does not require an inversion to obtain \( \hat{G}_{jk} \) and \( \hat{g}_{jk} \). Instead, the
mean \( \hat{g}_{jk} \) can be found by solving \( \hat{G}_{jk}^{-1} \hat{g}_{jk} = \tau_{jk}^2 \mathbf{K}_{jk} \hat{g}_{jk0} + \mathbf{P}_{jk}' \mathbf{V}_{jk}^{-1} \xi_{jk} \) by back substitution in \( O(n) \) operations. A random draw from \( \mathcal{N}(\hat{g}_{jk}, \hat{G}_{jk}) \) can then be efficiently obtained by sampling \( \nu \sim \mathcal{N}(0, \mathbf{I}) \), and solving \( \mathbf{C}z = \nu \) for \( z \) by back substitution, where \( \mathbf{C} \) is the Cholesky decomposition of \( \hat{G}_{jk}^{-1} \) and is also banded. It follows that \( z \sim \mathcal{N}(0, \hat{G}_{jk}) \), so that adding the mean \( \hat{g}_{jk} \) to \( z \) gives a draw \( g_{jk} \sim \mathcal{N}(\hat{g}_{jk}, \hat{G}_{jk}) \).

**Sampling** \( \tau_{jk}^2 \) The smoothness parameters \( \tau_{jk}^2 \) for each unknown function \( (j = 1, \ldots, J, k = 1, \ldots, q_j) \) are sampled from

\[
\tau_{jk}^2 | \theta, \tau_{jk}^2 \sim \mathcal{IG} \left( \frac{\nu_{jk0} + p_{jk} - 1}{2}, \frac{\delta_{jk0} + (\hat{g}_{jk} - \hat{g}_{jk0})' \mathbf{K}_{jk} (\hat{g}_{jk} - \hat{g}_{jk0})}{2} \right).
\]

**Sampling** \( y_{i4}^* \) Following Chib (1992), this full conditional density is seen to be truncated normal:

\[
y_{i4}^* | y, \theta \sim T \mathcal{N}_{(-\infty, 0)}(\mathbf{x}_{i4}' \beta_4 + \mathbf{w}_{i4} + E(\varepsilon_{i4} | \varepsilon_{i4}), \text{Var}(\varepsilon_{i4} | \varepsilon_{i4})), \quad i \in N_2.
\]

### 3.2 Modifications for multiple qualitative variables

If there is more than one qualitative variable in the model—e.g., the response variable or one or more of the endogenous variables is binary—three modifications to the basic scheme are required, following the framework of Albert and Chib (1993). First, \( y_{ij}^* \) is substituted for \( y_{ij} \) in the specification of the likelihood function, analogously to the use of \( y_{i4}^* \) in the selection equation. Second, \( y_{ij}^* \) is added to the sampler and sampled from an appropriately truncated distribution. Third, the variances of any binary or ordinal variables are set to one. If the response variable or any of the endogenous variables are censored, they are treated like \( y_{i4} \) in the discussion above with no variance restrictions.

The presence of binary or qualitative variables requires a modification to the algorithm to reflect the unit-variance constraints. When only one variable is binary or ordinal and that variable is always observed, the method for sampling \( \Omega \) presented in Section 3.1 may be utilized repeatedly to produce a draw for \( \Omega \). For instance, if \( y_{i4} \) is binary, rather than censored, the matrix \( \Omega_{22} \) in equation (7) is

\[
\Omega_{22} = \begin{pmatrix} \omega_{33} & \omega_{34} \\ \omega_{34} & 1 \end{pmatrix},
\]

so that the full conditional distributions for \( \omega_{34} \) and \( \omega_{33,4} = \omega_{33} - \omega_{34}^2 \) are normal and inverse Wishart (or inverse gamma when \( \omega_{33,4} \) is univariate), respectively. Notice that if
ω_{33} = 1 and ω_{44} is free, an obvious reindexing would again permit Gibbs sampling from inverse Wishart (or inverse gamma) and Gaussian distributions. Once Ω_{22} is obtained, the rest of Ω can be simulated in a straightforward way with the sampler in Section 3.1. A sampler for Ω in which one of the variances is restricted to unity is presented in Munkin and Trivedi (2003) in a setting with endogeneity but without incidental truncation.

When more than one variable is qualitative or when a unit restriction appears in Ω_{11}, i.e., there is a qualitative variable in y_{1:2}, the above derivations leading to normal and inverse Wishart sampling steps cannot be applied. Instead, the possibly multiple unit-variance and positive definiteness restrictions on Ω require a Metropolis-Hastings algorithm as in Chib and Greenberg (1998).

It is difficult to offer guidance on how such complications affect computational time. The key burden is in the sampling of the free elements of the restricted Ω. Although there are no costs of sampling such fixed elements of Ω as unit variances or zero correlations, and the computational costs can be quite small if Ω is well-structured and parameterized parsimoniously (e.g., equicorrelated or Toeplitz), sampling of the free elements requires a M-H step for which the implementation cost depends on whether tailored approximations to the full-conditional for Ω can be easily obtained by Taylor series methods or otherwise.

4 Model Comparison

Bayesian model comparison based on posterior odds ratios or Bayes factors requires computation of the marginal likelihood for each model under consideration. The marginal likelihood of our model is given by the integral

\[ m(y) = \int f(y|\beta, \Omega, \{g_{jk}\}, \{\tau_{jk}^{2}\})\pi(\beta, \Omega, \{g_{jk}\}, \{\tau_{jk}^{2}\}) d\beta d\Omega d\{g_{jk}\} d\{\tau_{jk}^{2}\}. \]

We compute the marginal likelihood by the approach of Chib (1995), who points out that the multivariate integral can be estimated from the identity

\[ m(y) = \frac{f(y|\beta^{*}, \Omega^{*}, \{g_{jk}^{*}\}, \{\tau_{jk}^{2*}\})\pi(\beta^{*}, \Omega^{*}, \{g_{jk}^{*}\}, \{\tau_{jk}^{2*}\})}{\pi(\beta^{*}, \Omega^{*}, \{g_{jk}^{*}\}, \{\tau_{jk}^{2*}\} | y)}, \]

where \(\beta^{*}, \Omega^{*}, \{g_{jk}^{*}\}, \) and \(\{\tau_{jk}^{2*}\}\) are fixed at high-density values. The posterior ordinate in the denominator of (16) must be estimated, but the likelihood and the prior ordinates in the
numerator are directly available. To estimate the denominator we use the decomposition

\[ \pi(\beta^*, \Omega^*, \{g_{jk}^*\}, \{\tau_{jk}^{2*}\} | y) = \]

\[ = \pi(\Omega^*, \{\tau_{jk}^{2*}\} | y) \pi(\beta^* | y, \Omega^*, \{\tau_{jk}^{2*}\}) \prod_{I_{mn}=1}^q \pi\left( g_{I_{mn}}^* | y, \Omega^*, \beta^*, \{\tau_{jk}^{2*}\}, \{g_{I_{jk}}^*\}_{I_{jk} < I_{mn}} \right), \]

where \( I_{jk} \) denotes a particular indexing of the \( q \) functions in the set \( \{g_{jk}\} \), so that it is not necessary to estimate the ordinates for each function in the order in which the functions appear in the model. The terms in the product are estimated by Rao-Blackwellization by averaging the full conditional densities of Section 3 with respect to MCMC draws coming from appropriately structured MCMC runs. In particular, the first ordinate is estimated with draws from the main MCMC run, whereas the remaining ordinates in the product are evaluated with MCMC output from reduced runs in which the parameters whose ordinates have already been obtained are held fixed and sampling is over the remaining elements of \( \theta \) and the latent data \( \{y_{14}^s\} \):

\[ \pi(\theta^*_s | y, \{\theta^*_j\}_{j<s}) = \frac{1}{G} \sum_{g=1}^G \pi(\theta^*_s | y, y_{14}', \{\theta^*_j\}_{j<s}, \{\theta_{j,g}^*(\theta)\}_{j>s}), \]

for draws \( \{\theta_{j,g}^*(\theta)\}_{j>s} \sim \pi\left( \{\theta_{j,g}^*(\theta)\}_{j>s} | y, \{\theta^*_j\}_{j<s} \right), g = 1, \ldots, G. \)

For the estimation of the marginal likelihood for semiparametric models, we make the following remarks. First, the numerical standard error of the marginal likelihood estimate, which indicates the variation that can be expected if the simulation were to be repeated, can be calculated by the method in Chib (1995). Second, the choice of a suitable posterior density decomposition is very important in this model because it determines the balance between computational and statistical efficiency. The large dimension of \( \{g_{jk}\} \), which may exceed the sample size, may increase the variability in the Rao-Blackwellization step if the full-conditional densities for the nonparametric functions are averaged over a conditioning set that changes with every iteration. For this reason, these large-dimensional blocks should be placed towards the end of the decomposition so that more blocks in the conditioning set remain fixed. This strategy leads to higher statistical efficiency and comes at a reasonable computational cost since the sampling of the unknown functions is \( O(n) \). Finally, the ordinate \( \pi(\Omega^*, \{\tau_{jk}^{2*}\} | y) \) is estimated jointly, rather than in individual reduced runs, because the parameters \( \Omega \) and \( \{\tau_{jk}^{2*}\} \) are conditionally independent given \( \beta \) and \( \{g_{jk}\} \). This observation can significantly reduce the computations.
5 Simulation Study

To study the effectiveness of the sampler in estimating nonparametric functions we simulate data from a 3-equation version of the model in equations (1)–(4) that is partly motivated by our subsequent application to log-wages considered in Section 6. In our simulation study each equation contains an intercept and two additive functions:

\begin{align}
  y_{i1} &= \beta_1 + g_{11}(y_{i2}) + g_{12}(w_{i11}) + \varepsilon_{i1}, \\
  y_{i2} &= \beta_2 + g_{21}(w_{i21}) + g_{22}(w_{i22}) + \varepsilon_{i2}, \\
  y_{i3}^* &= \beta_3 + g_{31}(w_{i31}) + g_{32}(w_{i32}) + \varepsilon_{i3}.
\end{align}

The nonparametric functions, graphed in Figure 1, have previously appeared in the literature: $g_{11}(y_{i2}) = 2\Phi(v) - 1$, where $\Phi(\cdot)$ is the standard normal c.d.f.; $g_{12}(v) = -0.8 + v + \exp(-30(v - 0.5)^2)$ for $v \in [0, 1]$; $g_{21}(v) = 1.5(\sin(\pi v))^2$ for $v \in [0, 1]$; $g_{22}(v) = \sin(v) + 1.5 \exp(-10v^2)$ for $v \in [-2, 4]$; $g_{31}(v) = 6(1 - \cos((\pi v/4)^2))$ for $v \in [0, 1]$; and $g_{32}(v) = 6v^3(1 - v^3)$ for $v \in [0, 1]$.

We assume that $y_{i2}$ is observed for all $i$. Each function that depends on exogenous covariates is evaluated at $m = 51$ equally spaced points on the support of each function, and $g_{11}(y_{i2})$ is evaluated at all points generated by the random realizations of $y_{i2}$. The data are generated with the equicorrelated covariance matrix $\Omega = (0.15I_4 + 0.1\mathbf{i}\mathbf{i}'$, which
implies a relatively high correlation of 0.4 between the errors in the individual equations.

Table 1 displays the signal-to-noise ratios for the functions, defined as the ratio of the range of the function to the standard deviation of the errors. These ratios vary from high noise (ratios around 2–3) to medium noise (ratios around 4–5). All else equal (e.g., the functional form, the sample size, and the number of observations at each design point), the functions tend to be estimated more precisely as the signal-to-noise ratio is increased (cf. Wood et al. 2002). This study covers a reasonably “strong noise” scenario, and the performance of the techniques is likely to improve as the sample size is increased, more points are introduced at each design point, or the error variances are decreased.

<table>
<thead>
<tr>
<th></th>
<th>Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Range($g_{k1}^j$)/SD($\varepsilon_{ij}$)</td>
<td>4.0 3.0 3.0 5.0 2.2 3.0</td>
</tr>
</tbody>
</table>

Table 1: Signal-to-noise ratios for the functions in the simulation study. (Note: the ratio for $g_{11}^j(\cdot)$ is an upper bound since the $y_{i2}$ are randomly generated and subsequently censored, thus unlikely to fill the entire range.)

To simplify the discussion of prior distributions, we denote the nonlinear functions by $g_k$, $k = 1, \ldots, 6$, in consecutive order over the three equations. Comparable priors are set for the equations: $\beta \sim \mathcal{N}(0, 5 \times I)$, $\Omega \sim IW(J + 4, 1.2 \times I_J)$, $g_k | \tau_k^2 \sim \mathcal{N}(0, \tau_k^2/E(\tau_k^2))$, $\tau_1^2 \sim IG(6, .0004)$, and $\tau_k^2 \sim IG(6, .04)$ for $k = 2, \ldots, 6$; these priors imply that $E(\Omega) = 0.4 \times I$, $SD(diag(\Omega)) = 0.57 \times 1$, $E(\tau_1^2) = SD(\tau_1^2) = 0.0001$, $E(\tau_k^2) = SD(\tau_k^2) = 0.01$ for $k = 2, \ldots, 6$, and $E_{\tau_k^2}(Var(g_k|\tau_k^2)) = 1$, for $k = 1, \ldots, 6$. A tighter prior on the smoothness parameter is assumed for the first function because most of the generated $y_{i2}$ appear around the middle of the support, where the function is approximately linear. Of course, the data play a role along with the prior in determining the posterior distribution of $\tau_1^2$. As is well known in the literature, high values of $\tau^2$ lead to undersmoothing as the function becomes less smooth and tries to interpolate the observations, whereas low values lead to smoother functions. Values of $\tau^2$ that are too high or too low yield poorer approximations that tend to improve as more data become available. An example of the role of the smoothness parameter in over- and undersmoothing is presented in Chib and Jeliazkov (2006).

We summarize and report the performance of the sampler over 20 Monte Carlo replications for $n = 500, 1000, 2000$. Due to randomness in obtaining the selected sample, $n_1$
varies across the simulations: under the simulation design, the selected sample $n_1$ comprises approximately 85% of the potential sample $n$ and varies from a low of 81.6% to a high of 86.6%. In comparison, our log-wage application in Section 6 has $n = 753$ and $n_1 = 428$, while in our simulated data, when $n = 500$, $n_1$ varies from 408 to 433. An important feature of the real data application compared to our simulation study is the presence of repeated values in the real data, which aids estimation. In our simulation study, most functions are specified to have $p = 51$ knots, except for $g_{11}(\cdot)$, which has $n_1$ knots with only one observation per knot, so that, in the first equation, the estimated $g_{11}$ and $g_{12}$ jointly contain more elements that the selected sample size $n_1$. Although the remaining equations require the estimation of lower-dimensional objects, the simulated data require more parameters than the real data application to log-wages, where all functions have fewer than 50 knots (ranging from 18 to 45), allowing more data to be available at each knot, which serves to estimate the functions more reliably than in the simulated data.

The posterior mean estimates $\hat{g}_{jk} = E(g_{jk}(v) | y)$, are found from MCMC runs of length 15000 following 2500 burn-in cycles. In the current context, the computational cost per 1000 MCMC draws is approximately 19 seconds when $n = 500$ and grows linearly with the sample size (approximately 38 seconds when $n = 1000$ and approximately 77 seconds when $n = 2000$), as is to be expected of an $O(n)$ algorithm. We gauge the performance of the method in fitting these functions by the root mean squared error, $\text{RMSE}_{jk} = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (\hat{g}_{jk}(v_{ijk}) - g_{jk}(v_{ijk}))^2}$, where the true functions are shifted so as to satisfy the identification constraints of the model. Boxplots of the $\text{RMSE}_{jk}$ for each function over the different samples are reported in Figure 2, where we see that the functions are estimated more precisely as the sample size grows. The function $g_{11}(y_{i2})$ is estimated less precisely than the other functions for two reasons: (1) all values in its design point vector are unique, whereas the other functions are evaluated at a smaller number of unique design point values, and (2) the first two functions are estimated from only $n_1$ observations, while the remaining functions are estimated from the full set of $n$ observations. Finally, a close examination of the model structure shows that the level of $g_{11}(\cdot)$ in (17) is related not only to the intercept in its own equation, as is generally true for all functions in additive models (see the discussion on identification in Section 2.2), but is also correlated by construction with parameters in $\Omega$. This can be seen from the description of the sampler for $g$, especially
equation (15), which shows that the errors in the endogenous covariate equation determine both the endogenous covariate and the conditional mean used in sampling $g_{11}(\cdot)$ through the covariance elements in $\Omega$. For the other functions, the errors in other equations determine the conditional mean, but are not related to any other features of the sampled function such as its design point vector, which is what is special about $g_{11}(\cdot)$ in this model. Plots of the true and estimated functions for $n = 1000$ are shown in Figure 3. Both Figures 2 and 3 show that the method recovers the true functions well.

We compute the inefficiency factors resulting from the Markov chain for the parametric components of the model. The inefficiency factor is defined as $1 + 2 \sum_{l=1}^{L} \rho_k(l) (1 - l/L)$, where $\rho_k(l)$ is the sample autocorrelation at lag $l$ for the $k$th parameter in the sampling and the summation is truncated at values $L$ at which the correlations taper off. This quantity may be interpreted as the ratio of the numerical variance of the posterior mean from the MCMC chain to the variance of the posterior mean from hypothetical independent draws. Figure 4, which displays the inefficiency factors obtained from the sampler discussed in Section 3.1, suggests several conclusions. First, although the inefficiency factors for $\beta$ are the largest and do not seem to depend on the sample size, they are well within the limits found in the MCMC literature dealing with similar models (e.g., Chib and Greenberg 2007). For this reason, we suggest that a longer MCMC chain may be required for more accurate estimation of $\beta$, but there are no other adverse consequences. Second, the elements of $\Omega$ are sampled very efficiently (some are i.i.d.), and the parameters of $\Omega$ that enter the Tobit equation (the last three elements of $\text{vech}(\Omega) = (\omega_{11}, \omega_{21}, \omega_{22}, \omega_{31}, \omega_{32}, \omega_{33})$) are estimated

Figure 2: Boxplots of root mean squared errors of function estimates in the simulation study.
better as sample sizes increase because there are more latent data points. Finally, since the estimates of the $\{\tau_{jk}^2\}$ depend on how well the corresponding functions are sampled, they depend on the sample size in a predictable way: as the sample size grows and the $\{g_{jk}\}$ are estimated better, so are the corresponding $\{\tau_{jk}^2\}$.

We conclude the discussion of the simulated data by revisiting the discussion, below Algorithm 1 of Section 3.1, of the advantages of not involving the outcomes that are missing due to the selection mechanism. The inefficiency factors for our algorithm, which does not augment the sampler with the missing outcomes, are summarized in Figure 4. We next compare the performance of our algorithm to that of an algorithm that includes the missing outcomes. Specifically, samples are obtained from an algorithm in which $\{g_{jk}\}, \{\tau_{jk}^2\}, \beta,$ and $\{y_{i4}\}$ are sampled as before (without involving the missing data for the incidentally truncated outcomes). To sample $\Omega$, however, we first augment $\eta_{i2:3}$ for $i \in N_2$ with the missing $\eta_{i1}$ (see the sampler for $\Omega$ in Section 3.1, but keep in mind that here we are fitting a 3-equation system), which are drawn conditional on $\eta_{i2:3}$ and all other parameters and
Figure 4: Inefficiency factors for the parameters of the model in the simulation study.

data, including $\Omega$. After the data are “balanced” in this way, $\Omega$ can is sampled directly from an inverse Wishart distribution.

The results from the sampler augmented with the outcomes that are missing due to the selection mechanism are presented in Figure 5. A comparison of Figures 4 and 5 shows that inclusion of the missing outcomes results in drastic deterioration of the inefficiency factors. This result is most obvious in the inefficiency factors for the first equation, where augmentation is performed to “balance” the sample, and all sample sizes: for a sample of 500, the inefficiency factor without latent data for $\beta_1$ has a median of about 40, whereas its median is close to 200 when latent data are included. As another example, for $n = 2000$, the median is about 50 without the missing outcomes and 400 with the missing outcomes included in the sampling. Again, with $n = 2000$, the median inefficiency factor for $\omega_{12}$ is
about 2.5 without the missing outcomes and about 10 with them. We make several observations regarding these results. First, the sampler we introduced in Section 3 dominates a sampler augmented with the missing outcomes on all margins: the additional computational and storage demands of introducing the missing outcomes related to augmenting the sample with missing data do not pay off as the inefficiency factors are significantly higher. Second, Figure 5 reveals that the inefficiency factors actually rise in larger samples. This may seem surprising because we expect samplers to improve as more data become available. But an important feature of this model is that larger potential samples imply both larger selected samples and larger amounts of missing data to simulate, which slows the convergence of the Markov chain. Finally, note that we generate auxiliary data only in the sampling of $\Omega$, while all other parameters are sampled marginally of it. If one were to try an even more “naive” approach, where the missing data are involved in all steps of the sampler, the results would be even worse than those we have reported for the reasons we have mentioned. These conclusions are in close agreement with the results in Liu (1994) and Liu et al. (1994), who emphasize the importance of constructing Markov chains that sample parameters jointly instead of conditionally on each other.

6 Estimating Women’s Wages

We apply the techniques of this paper to study the determinants of women’s wages, a topic that has been extensively studied because of the large increases in women’s participation in the labor force and in hours of work in the U.S. in the postwar period. Goldin (1989) reports a seven-fold increase in participation of married women since the 1920s, and Heckman (1993) underscores the importance of participation (entry and exit) decisions in estimating labor supply elasticities. The empirical analysis of women’s labor supply is complicated by the possible endogeneity of a covariate and by sample selection concerns. Endogeneity may be a concern for the level of education, a covariate that affects wages, since education is likely to be affected by such unobserved variables as motivation, work ethic, perseverance, and intelligence, that are also determinants of wage rates. The problem of sample selection arises because wages are not observed for women who report zero annual hours of work, and such women may be out of the labor force because of excessively low wage offers. Our model allows for sample selection and endogeneity, and the nonparametric functions
allow us to explore the presence of nonlinearities in the effects of some of the covariates. Such nonlinearities have been modeled by the inclusion of quadratic terms for certain variables; see Mroz (1987) and Wooldridge (2002). We extend these results by estimating a semiparametric specification and comparing it to a number of alternatives models.

The data set, which is available on the JCGS website, is from Mroz (1987). The potential sample consists of 753 married women, 428 of whom are employed and therefore in the selected sample. The variables in the data set are summarized in Table 2. We specify the
Table 2: Variables in the women’s labor supply example from Mroz (1987). The sample consists of 753 married women, 428 of whom work. All summary statistics are for the full sample except where indicated.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Explanation</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>WAGE</td>
<td>woman’s wage rate (only for those working)</td>
<td>4.18</td>
<td>3.31</td>
</tr>
<tr>
<td>EDU</td>
<td>woman’s years of schooling</td>
<td>12.29</td>
<td>2.28</td>
</tr>
<tr>
<td>HRS</td>
<td>woman’s hours of work in 1975</td>
<td>740.58</td>
<td>871.31</td>
</tr>
<tr>
<td>AGE</td>
<td>woman’s age in years</td>
<td>42.54</td>
<td>8.07</td>
</tr>
<tr>
<td>EXPER</td>
<td>actual labor market experience in years</td>
<td>10.63</td>
<td>8.07</td>
</tr>
<tr>
<td>KLT6</td>
<td>number of kids under 6 years old</td>
<td>0.28</td>
<td>0.52</td>
</tr>
<tr>
<td>KGE6</td>
<td>number of kids 6–18 years old</td>
<td>1.35</td>
<td>1.32</td>
</tr>
<tr>
<td>NWINC</td>
<td>estimated nonwife income (1975, in $10,000)</td>
<td>2.01</td>
<td>1.16</td>
</tr>
<tr>
<td>MEDU</td>
<td>mother’s years of schooling</td>
<td>9.25</td>
<td>3.37</td>
</tr>
<tr>
<td>FEDU</td>
<td>father’s years of schooling</td>
<td>8.81</td>
<td>3.57</td>
</tr>
<tr>
<td>HEDU</td>
<td>husband’s years of schooling</td>
<td>12.49</td>
<td>3.02</td>
</tr>
</tbody>
</table>

The econometric model as

\[
\begin{align*}
  y_{i1} &= x_{i1}'\beta_1 + g_{i1}(y_{i2}) + g_{i2}(w_{i11}) + \varepsilon_{i1}, \\
  y_{i2} &= x_{i2}'\beta_2 + g_{i2}(w_{i21}) + g_{i22}(w_{i22}) + \varepsilon_{i2}, \\
  y_{i3}^* &= x_{i3}'\beta_3 + g_{i3}(w_{i31}) + g_{i32}(w_{i32}) + \varepsilon_{i3},
\end{align*}
\]

where \( y_{i3} = y_{i3}^* I(y_{i3}^* > 0) \) is the Tobit selection variable, so that \( y_{i1} \) is observed only when \( y_{i3} \) is positive, and \( y_{i2} \) is always observed. Based on the parametric models discussed in Mroz (1987) and Wooldridge (2002), we let

\[
\begin{align*}
  y_i &= (y_{i1}, y_{i2}, y_{i3})' = (\ln(WAGE_i), EDU_i, \sqrt{HRS_i})', \\
  x_{i1} &= 1, \\
  x_{i2} &= x_{i3} = (1, KLT6_i, KGE6_i, NWINC_i, MEDU_i, FEDU_i, HEDU_i)', \\
  w_{i21} &= w_{i31} = AGE_i, \quad \text{and} \quad w_{i11} = w_{i22} = w_{i32} = EXPER_i.
\end{align*}
\]

Note that this model is a slightly simplified version of the example set out in Section 1 because it does not include on-the-job training, which is an endogenous variable available only for the selected sample. The choice of covariates and instruments in our model is consistent with earlier studies, but differs in the way in which the covariates are allowed to affect the responses. The nonparametric specification for \( AGE_i \) and \( EXPER_i \) is of particular interest because these covariates embody cohort, productivity, and life-cycle effects that are likely to affect wages nonlinearly. Wooldridge (2002, chap. 17) considers parametric models
that contain linear and quadratic terms in EXPER. The parameter estimates for our model are given in Table 3, and the nonparametric functions are plotted in Figure 6.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Covariate</th>
<th>Mean</th>
<th>SD</th>
<th>Median</th>
<th>Lower</th>
<th>Upper</th>
<th>Ineff</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>1</td>
<td>0.113</td>
<td>0.360</td>
<td>0.103</td>
<td>-0.558</td>
<td>0.806</td>
<td>43.930</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>1</td>
<td>4.817</td>
<td>0.395</td>
<td>4.821</td>
<td>4.037</td>
<td>5.597</td>
<td>12.388</td>
</tr>
<tr>
<td></td>
<td>KLT6</td>
<td>0.229</td>
<td>0.131</td>
<td>0.228</td>
<td>-0.027</td>
<td>0.485</td>
<td>3.762</td>
</tr>
<tr>
<td></td>
<td>KGE6</td>
<td>-0.084</td>
<td>0.056</td>
<td>-0.084</td>
<td>-0.193</td>
<td>0.024</td>
<td>2.172</td>
</tr>
<tr>
<td></td>
<td>NWINC</td>
<td>0.144</td>
<td>0.059</td>
<td>0.145</td>
<td>0.030</td>
<td>0.259</td>
<td>1.699</td>
</tr>
<tr>
<td></td>
<td>MEDU</td>
<td>0.134</td>
<td>0.023</td>
<td>0.134</td>
<td>0.090</td>
<td>0.178</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>FEDU</td>
<td>0.093</td>
<td>0.021</td>
<td>0.094</td>
<td>0.051</td>
<td>0.135</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>HEDU</td>
<td>0.347</td>
<td>0.023</td>
<td>0.347</td>
<td>0.301</td>
<td>0.393</td>
<td>1.421</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>1</td>
<td>-0.493</td>
<td>2.101</td>
<td>-0.501</td>
<td>-4.611</td>
<td>3.570</td>
<td>1.914</td>
</tr>
<tr>
<td></td>
<td>KGE6</td>
<td>-0.037</td>
<td>0.760</td>
<td>-0.036</td>
<td>-1.537</td>
<td>1.450</td>
<td>1.712</td>
</tr>
<tr>
<td></td>
<td>NWINC</td>
<td>-1.197</td>
<td>0.924</td>
<td>-1.194</td>
<td>-3.019</td>
<td>0.591</td>
<td>2.230</td>
</tr>
<tr>
<td></td>
<td>MEDU</td>
<td>0.480</td>
<td>0.354</td>
<td>0.479</td>
<td>-0.211</td>
<td>1.170</td>
<td>1.336</td>
</tr>
<tr>
<td></td>
<td>FEDU</td>
<td>0.164</td>
<td>0.345</td>
<td>0.166</td>
<td>-0.520</td>
<td>0.845</td>
<td>1.262</td>
</tr>
<tr>
<td></td>
<td>HEDU</td>
<td>-0.028</td>
<td>0.330</td>
<td>-0.029</td>
<td>-0.672</td>
<td>0.620</td>
<td>4.104</td>
</tr>
<tr>
<td>$\omega_{11}$</td>
<td></td>
<td>0.441</td>
<td>0.032</td>
<td>0.440</td>
<td>0.383</td>
<td>0.507</td>
<td>2.053</td>
</tr>
<tr>
<td>$\omega_{21}$</td>
<td></td>
<td>0.112</td>
<td>0.081</td>
<td>0.112</td>
<td>-0.047</td>
<td>0.274</td>
<td>10.969</td>
</tr>
<tr>
<td>$\omega_{22}$</td>
<td></td>
<td>2.752</td>
<td>0.143</td>
<td>2.747</td>
<td>2.484</td>
<td>3.046</td>
<td>1.000</td>
</tr>
<tr>
<td>$\omega_{31}$</td>
<td></td>
<td>-0.625</td>
<td>1.593</td>
<td>-0.610</td>
<td>-3.786</td>
<td>2.467</td>
<td>7.488</td>
</tr>
<tr>
<td>$\omega_{32}$</td>
<td></td>
<td>6.262</td>
<td>1.685</td>
<td>6.245</td>
<td>3.031</td>
<td>9.667</td>
<td>1.346</td>
</tr>
<tr>
<td>$\omega_{33}$</td>
<td></td>
<td>632.960</td>
<td>47.749</td>
<td>630.570</td>
<td>547.135</td>
<td>734.045</td>
<td>3.212</td>
</tr>
<tr>
<td>$\tau_1^2$</td>
<td></td>
<td>0.008</td>
<td>0.005</td>
<td>0.006</td>
<td>0.003</td>
<td>0.020</td>
<td>4.176</td>
</tr>
<tr>
<td>$\tau_2^2$</td>
<td></td>
<td>0.004</td>
<td>0.002</td>
<td>0.004</td>
<td>0.002</td>
<td>0.009</td>
<td>6.433</td>
</tr>
<tr>
<td>$\tau_3^2$</td>
<td></td>
<td>0.005</td>
<td>0.003</td>
<td>0.004</td>
<td>0.002</td>
<td>0.012</td>
<td>6.083</td>
</tr>
<tr>
<td>$\tau_4^2$</td>
<td></td>
<td>0.005</td>
<td>0.002</td>
<td>0.004</td>
<td>0.002</td>
<td>0.011</td>
<td>6.598</td>
</tr>
<tr>
<td>$\tau_5^2$</td>
<td></td>
<td>0.011</td>
<td>0.009</td>
<td>0.008</td>
<td>0.003</td>
<td>0.035</td>
<td>11.550</td>
</tr>
<tr>
<td>$\tau_6^2$</td>
<td></td>
<td>0.029</td>
<td>0.023</td>
<td>0.023</td>
<td>0.007</td>
<td>0.089</td>
<td>15.585</td>
</tr>
</tbody>
</table>

Table 3: Parameter estimates for nonparametric model of women's wage function model under the priors $\beta \sim N(0, 5 \times I)$, $\Omega \sim IW(7, 1.2 \times I)$, $g_{2k} \tau_k^2 \sim N(0, \tau_k^2 / E(\tau_k^2))$, and $\tau_k^2 \sim IG(6, .04)$ for $k = 1, \ldots, 6$. The table also reports 95% credibility intervals and inefficiency factors from 25000 MCMC iterations.

The estimates in Table 3 are consistent with the predictions of economic theory. Results of the education equation reveal that the presence of younger children is associated with a higher level of mother's education than having older children; presumably, having children earlier in life interferes with a woman's education. The results also show that women who live in families with higher non-wife income, as well as women whose parents and
husband are better educated, are more likely to be better educated themselves. Results of the hours-worked equation suggest that having young children reduces the hours worked as evidenced by the negative mean and a 95% credibility interval that lies below zero, but older children have little impact on hours. Again, consistent with economic theory, higher non-wife income and lower parents’ schooling reduce hours of work. The effect of husband’s education is weak, both statistically and economically: its 95% credibility interval includes both negative and positive values, and its mean is small relative to its standard deviation.

The estimates of $\Omega$ provide evidence that education is endogenous: the 95% credibility interval of $\omega_{21}$ is mostly in positive territory. In addition, although the errors in the log wage equation are largely uncorrelated with those in the hours equation, sample selection is non-ignorable because the correlation between the errors in the education and hours equations is clearly positive.

We now consider the nonparametric functions plotted in Figure 6. These suggest that log-wages generally increase with education and experience, but that the increase is stronger for women with at least some college (the slope of $g_{11}$ appears to change around 14 years of schooling). Moreover, the first 7–8 years of job experience lead to rapid gains in wages, after which wages appear to stabilize. An interesting nonlinearity appears at the end of the range of experience, where women with over 30 years of experience appear to command high wage rates. The amount of schooling does not vary with age for women between 30 and 60 years old, when most people have completed school, but appears to be positively related to experience. Finally, the figure shows a strong negative effect of age on hours of work, which is consistent with cohort and life-cycle effects, and a strong positive effect of experience on hours, which is consistent with increases in productivity as experience grows.

Since, with a few exceptions, the nonparametric profiles do not show substantial curvature, we compare this model to several simpler alternatives. The log marginal likelihood of the model with six nonparametric functions is estimated to be $-4236.44$ with a numerical standard error of 0.144, and a model that models only $g_{11}(y_{12})$ nonparametrically, the other covariates being entered linearly, has an almost identical log marginal likelihood of $-4236.43$ with a numerical standard error of 0.075. Since the data do not provide enough information to distinguish between them, these two models appear equiprobable. These models are also compared to two parametric models—the log marginal likelihood estimate
for a linear model is $-4237.405$ with numerical standard error of 0.012, and a parametric model that includes experience squared in all three equations has a log marginal likelihood estimate of $-4244.58$ with numerical standard error of 0.012. In this application it appears that linearity is a reasonable assumption for most of the covariate effects, but there is some evidence supporting the possibility that at least one and possibly two of the effects, namely those of education and experience, are nonlinear.

Figure 6: Function estimates in the log-wage application.

7 Conclusions

This paper introduces an efficient approach to analyzing a general class of models in which the problem of sample selection arises. The models include linear and nonparametric components and may involve endogenous regressors that enter the response equation nonparametrically. The class of models may involve multi-equation systems of responses that comprise the selected sample or multi-equation systems that are always observed. The responses in these systems may be continuous, binary, ordered, or Tobit (censored).
An important aspect of our MCMC algorithm for this class of models is that it does not require simulation of the outcomes that are missing due to the selection mechanism, a feature of the estimation method that enhances computational efficiency as we show in our simulation experiments. Thus, one should not include such missing outcomes in the MCMC simulation when it is possible to proceed otherwise. In our case, even without the inclusion of these missing outcomes, all sampling is from full conditional distributions unless there are constraints on the covariance matrix arising from binary response or binary endogenous variables. In the latter cases, modified MCMC algorithms are available as we point out. The ability to compute marginal likelihoods makes it possible to compare different parametric and semiparametric model specifications in a fully Bayesian environment. A simulation study shows that the methods perform well, and an application involving a semiparametric model of women’s labor force participation and log-wage determination illustrates that the model and the estimation methods are practical and can uncover interesting features in the data.

References


34


